Capstone Project (Regression)

Ted Talk Views Prediction

By: Kumari Rashmi

Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Feature Extraction & Selection

D. Model Creation and Evaluation

E. Key Takeaways & Conclusion

What this presentation is about?

About TED Talks:

- Slogan “Ideas worth spreading”
- TED is a non-profit devoted to spreading ideas, usually in the form of short, powerful talks online (<18 minutes) on almost all topics, from science to business to global issues

Exploratory Data Analysis:

- Identifying Underlying Patterns and spotting Irregularities
- Drawing actionable insights
- Feature Engineering & Selection

Model Creation

- Building models to predict Views of a given TED Talk
- Evaluating models and choose best performing Model

Basic Information about Dataset

- There are total 4005 observations with 19 features, each talk is identified by unique Talk Id
- Out of 19 features, 4 features are numerical features, 2 are temporal features and rest are either categorical or textual features
- The average duration of a TED Talk is ~12 minutes
- The average comments and views for a TED talk are ~162 comments and ~2.1 millions
- Majority of TED Talks (~99%) are in English native language
- Alex Gendler (Speaker) has given the highest number of TED Talks (34 talks)

Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Feature Extraction & Selection

D. Model Creation and Evaluation

E. Key Takeaways & Conclusion

Data Preparation & Cleaning

1. Formatting Inconsistent data type:

- Date values of recorded and publishing columns
- List values of Available languages and topics
- Dictionary values of all_speakers, occupations and related_talks

2. Handling Missing Values:

- <0.1% values are missing in “all_speakers” & “recorded_date”; Since these are not relevant to our analysis, hence were ignored
- ~16% values are missing in “comments”. Accordingly missing comments were replaced by Median value of comments
- ~13% values are missing in “occupations”. Occupations has only been used to extract occupation of Main speaker. Accordingly, missing occupation for main speaker has been considered as “Other”

3. Handling Data Outliers:

a. Comments:

- Average and median comments are 162 and 89 respectively (highly skewed)
- Comments of Talks outside (+/-)1.5*IQR were capped

b. Duration:

- ~10% of TED talks have duration more than 19 minutes while average is ~12 minutes (highly skewed)
- Duration of Talks outside (+/-)1.5*IQR were capped

c. Views:

- Average and median views are ~2.2 millions and ~1.4 millions respectively (highly skewed)
- Views of Talks outside (+/-)1.5*IQR were capped

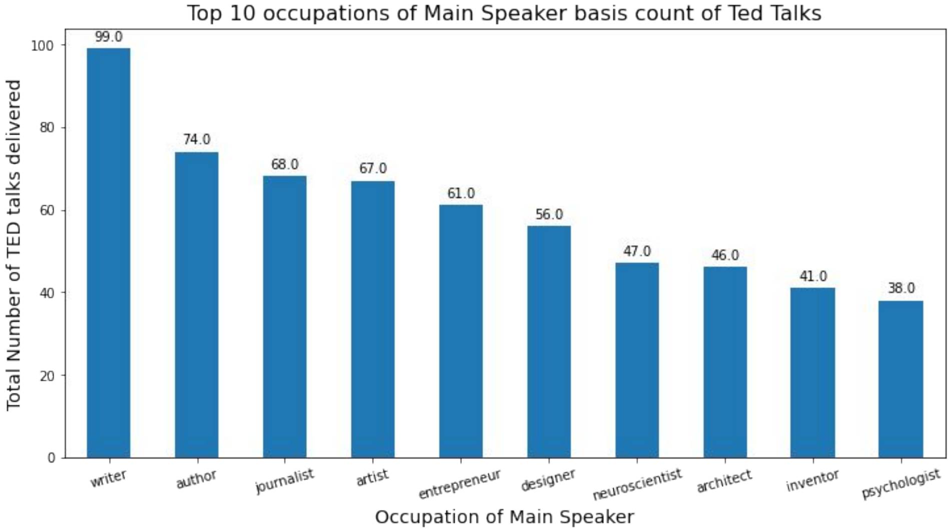
d. Number of Languages:

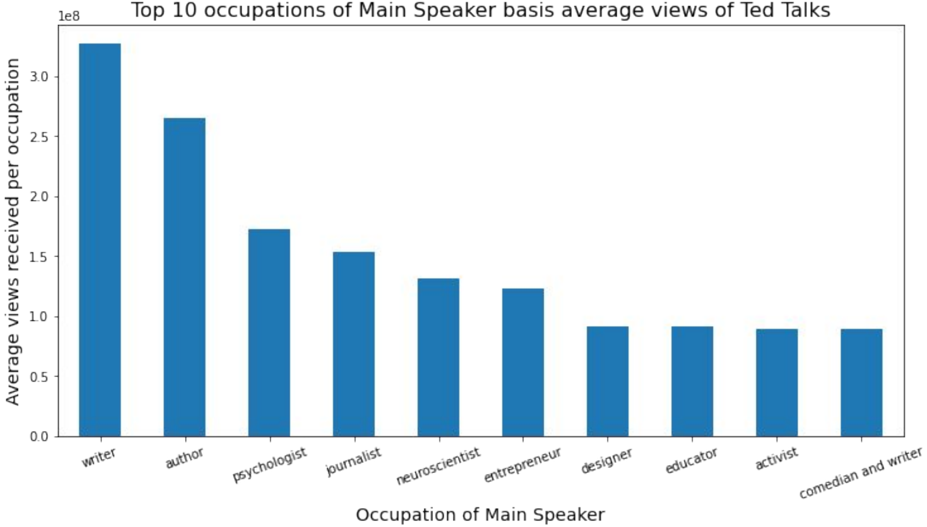
- Number of languages of Talks outside (+/-)1.5*IQR were capped

Underlying Patterns and Trends

1. Occupation of Main Speaker

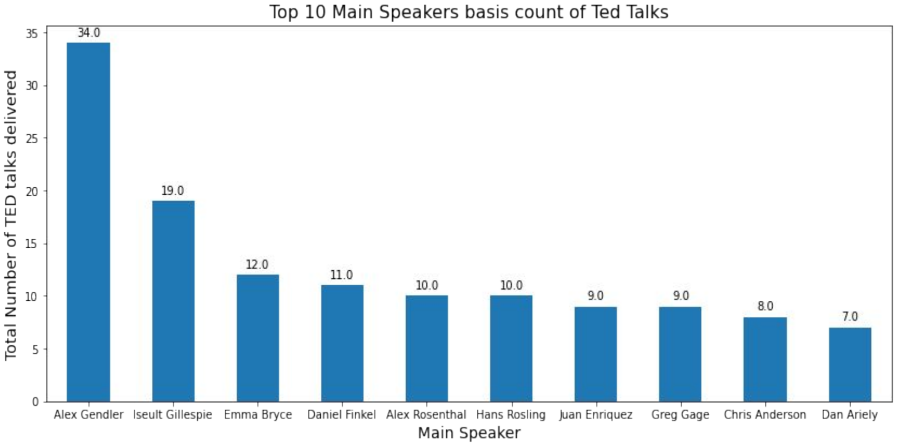
- Highest number of Ted Talks were delivered by Speakers who are writers, shortly followed by speakers who are author
- Average views received per TED talk is highest for Speakers who are writers, shortly followed by speakers who are author

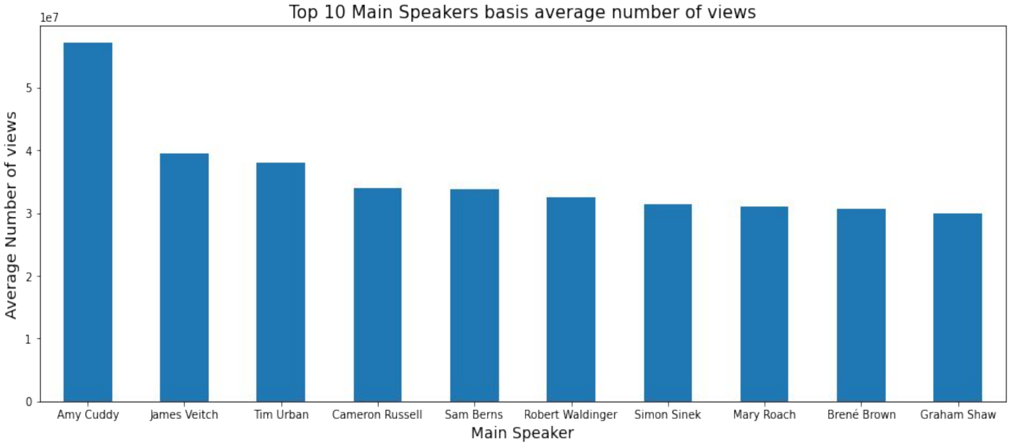

Top 10 occupations of Main Speaker basis count of Ted Talks


Top 10 occupations of Main Speaker basis average views of Ted Talks

2. Main Speakers

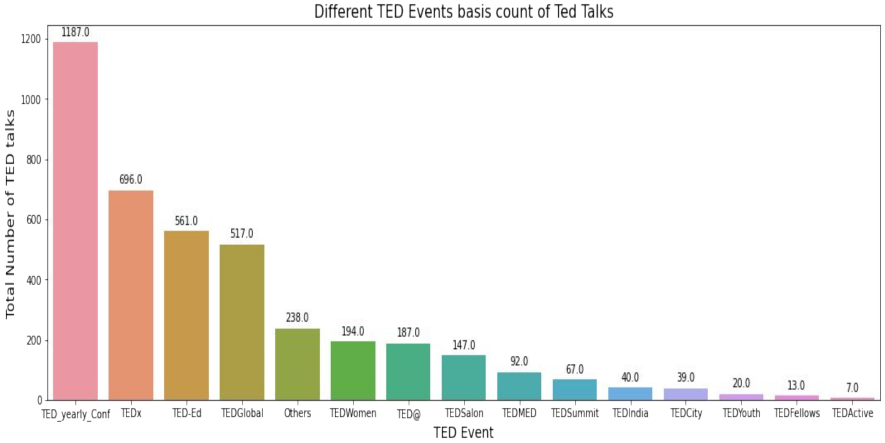
- Alex Gendler, educator of TED-Ed, has given the highest number of Ted Talks
- Amy Cuddy, a social psychologist, has the highest number of views per Talk
- Basis above, TED team may consider creating more talks with speakers with highest average views such as Amy Cuddy, James Veitch and Tim Urban

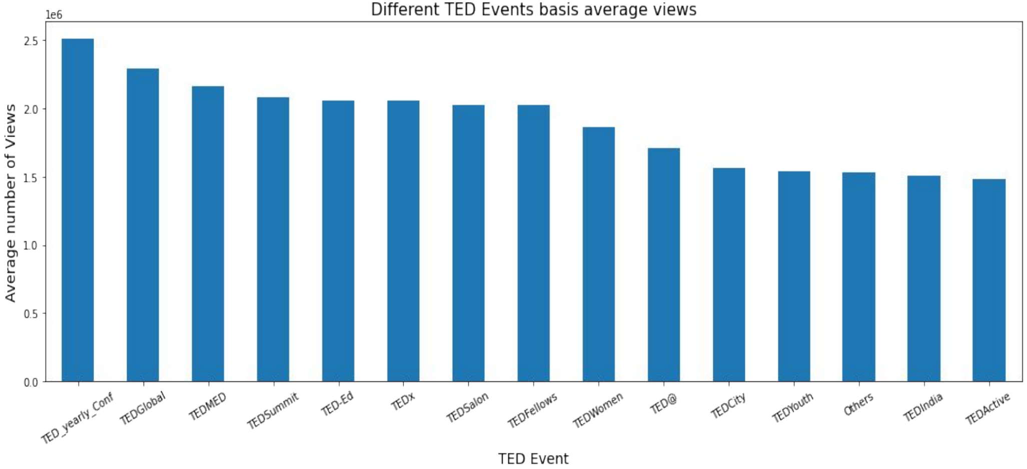




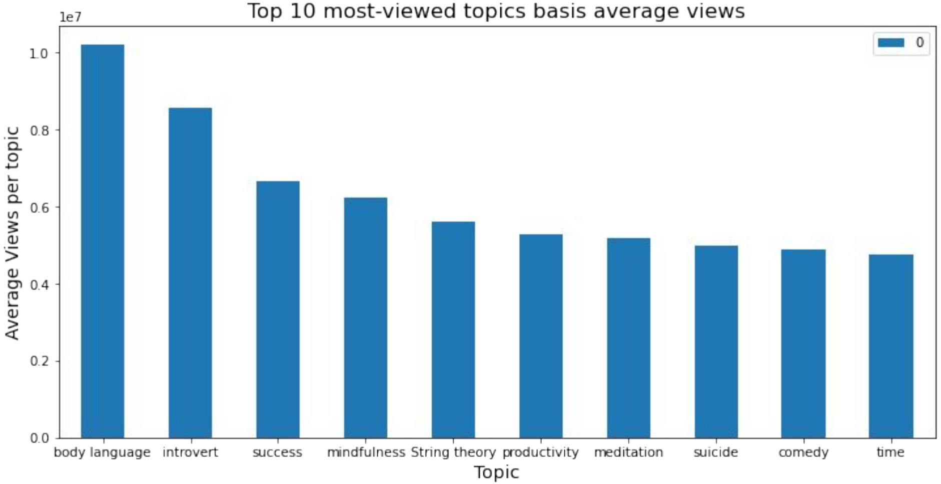
3. Different TED Events

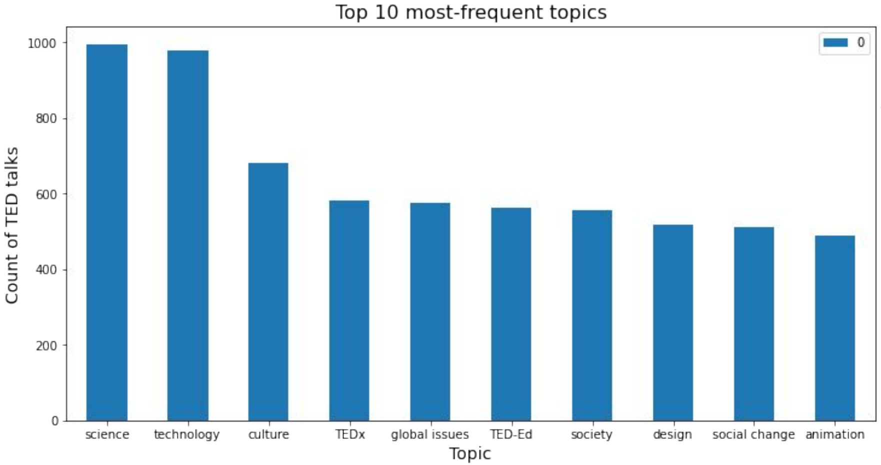
- TED Talks are organized under various events such as TED Conferences, TEDx, TED-Ed, TED Women, TED Global etc
- TED Conference has the highest number of talks; followed by TEDx & TED-ed
- “TED Conference” event has the highest average views per talk; shortly followed by “TED Global” event

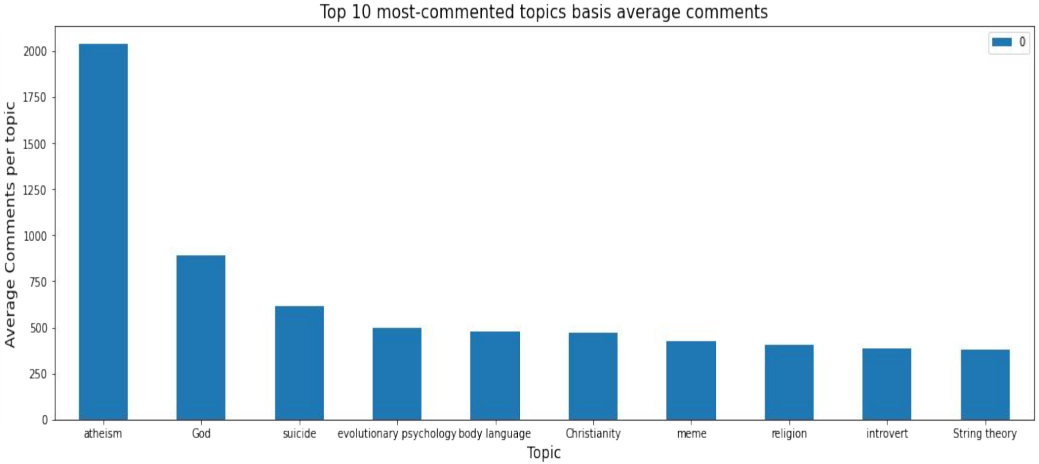




4. Most Frequent, Viewed & Commented Topics

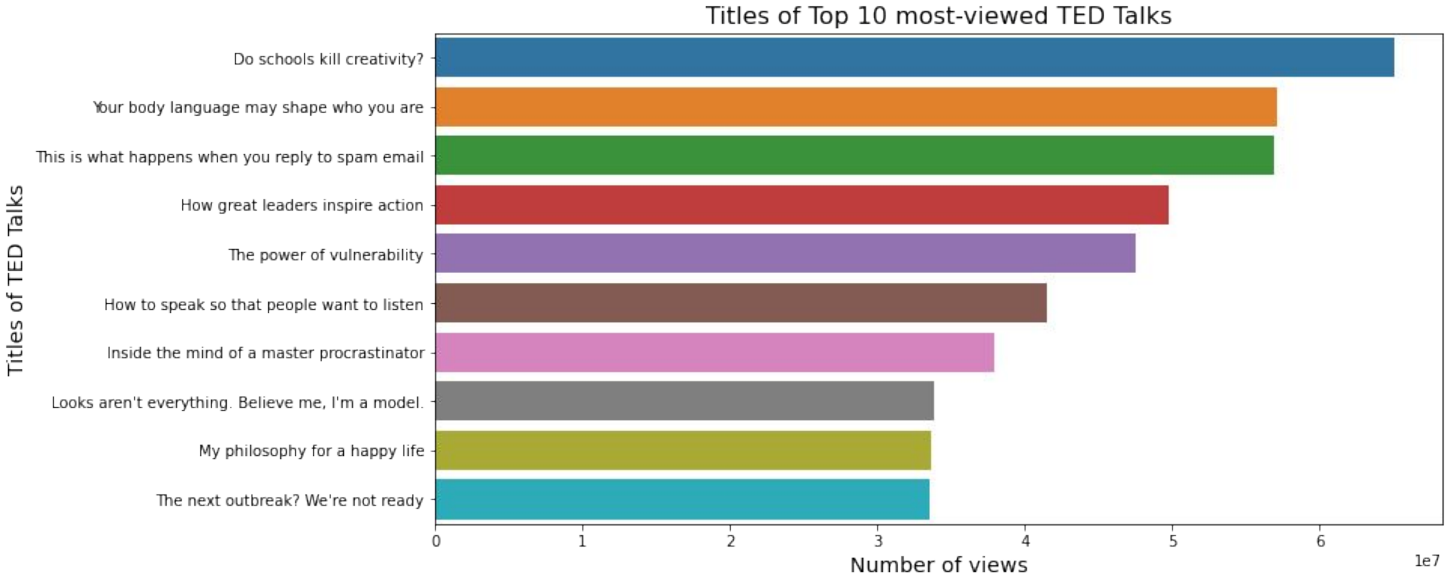






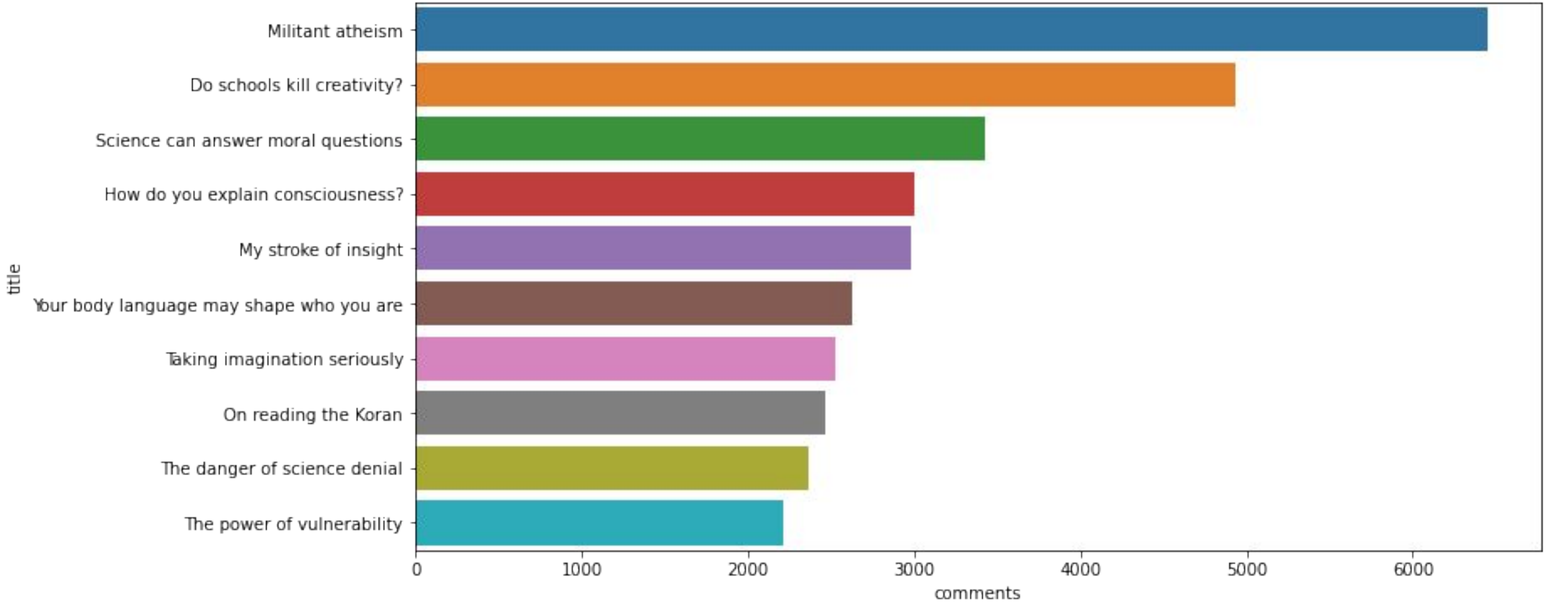
5. Top 10 Most-viewed TED Talks

- "Do schools kill creativity" by Sir Ken Robinson is the most viewed talk
- Top 10 most-viewed talks are either TEDx or TED Conference or TED Global events, and all are in English native language
- On average, Top 10 most-viewed talk is of ~14 minutes durations, has ~1452 comments and is available ~48 languages



6. Top 10 Most-commented TED Talks

- "Militant atheism" by Richard Dawkins is the most commented talk
- Most commented talks are based on topics as Atheism, creativity, science, psychology, culture, body language and Religion
- On average, top 10 most commented talk is of duration ~19 minutes, has ~3300 comments and is available in ~43 languages



7. Least-Viewed but Most-Commented Talks

- There are total 9 TED Talks, which have comments more than 85th percentile, but have views fewer than 15th percentile

- The average duration for above talks is ~12 minutes

- Above talks are based on topics such as Politics, Voting, Racism, Dance, Humanity, Internet, Nuclear Energy and Alternative Energy (Renewables)

- Since these talks are more commented but less viewed, it signals that these topics have potential to attract more audience and can be explored further

- While curating new TED Talks, TED team may explore above 9 talks and accordingly decide whether to create more similar contents around those topics or not

Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Feature Extraction & Selection

D. Model Creation and Evaluation

E. Key Takeaways & Conclusion

List of Features Extracted

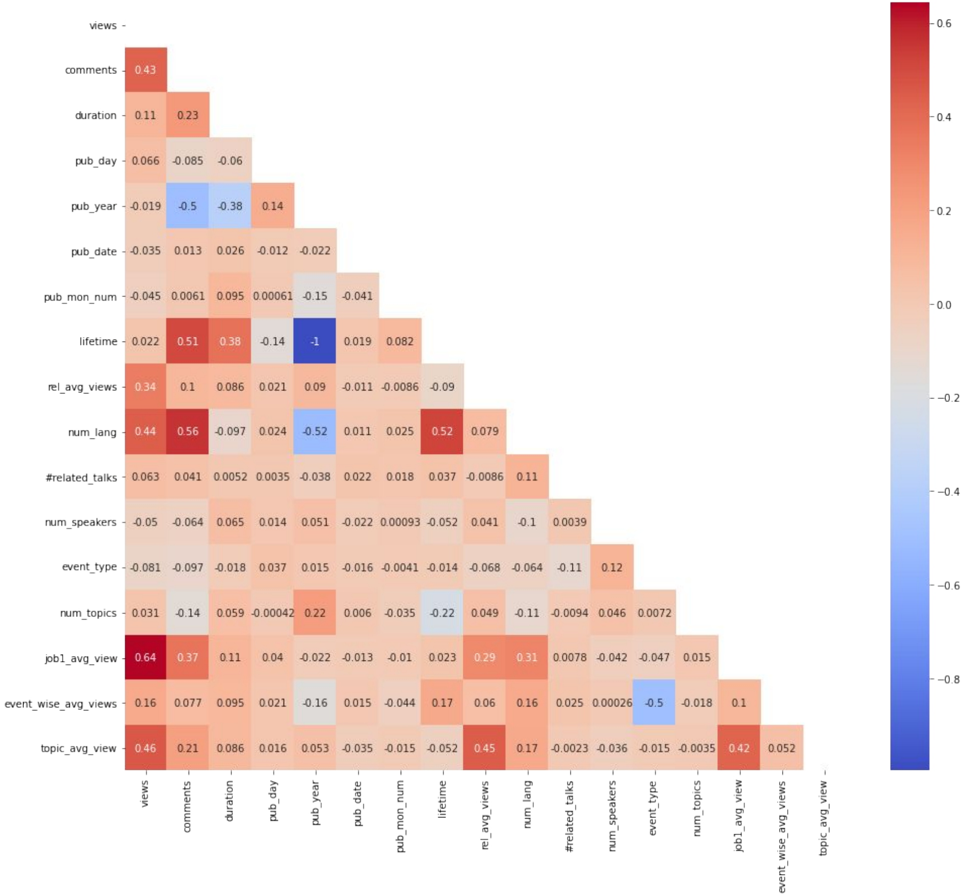
1. **Before Train-Test Splitting**:
   - Weekday, Date, Month and year of publishing
   - Lifetime, i.e., How many days ago a Talk was published
   - Number of languages in which a Talk is available
   - Number of Related Talks
   - Number of Speakers in each Talk (~97% Talks have 1 speaker only)
2. **After Train-Test Splitting** (To avoid Data Leakage):
   - Average views of all related talks for each ted talk
   - Average views for occupation of Main Speaker
   - Average views for all topics for each talk
   - Average views of event of each talk

Features Selection

In order to remove redundancy and avoid overfitting, followings were used to select the most informative features before modelling :

1. Correlation Matrix
2. SelectKBest class of sklearn library:
   - This filter method selects features according to the k highest scores basis F-statistics calculated for each input variable with the target
   - score_func used is “f_regression”
3. Sequential Forward Selection:
   - It is a wrapper method, where searching algorithm adds feature sequentially to an empty set of features until the addition of extra features does not reduce the criterion
   - RandomForestRegressor algorithm with “neg_mean_absolute_error” as criterion was used
4. Trends identified during EDA

Correlation Matrix



	views	comments	duration	pub_day	pub_year	pub_date	pub_mon_num	lifetime	rel_avg_views	num_lang	#related_talks	num_speakers	event_type	num_topics	job1_avg_view	event_wise_avg_views	topic_avg_view
views																	
comments	0.43																
duration	0.11	0.23															
pub_day	0.066	-0.085	-0.06														
pub_year	-0.019	-0.5	-0.38	0.14													
pub_date	-0.035	0.013	0.026	-0.012	-0.022												
pub_mon_num	-0.045	0.0061	0.095	0.00061	-0.15	-0.041											
lifetime	0.022	0.51	0.38	-0.14	-1	0.019	0.082										
rel_avg_views	0.34	0.1	0.086	0.021	0.09	-0.011	-0.0086	-0.09									
num_lang	0.44	0.56	-0.097	0.024	-0.52	0.011	0.025	0.52	0.079								
#related_talks	0.063	0.041	0.0052	0.0035	-0.038	0.022	0.018	0.037	-0.0086	0.11							
num_speakers	-0.05	-0.064	0.065	0.014	0.051	-0.022	0.00093	-0.052	0.041	-0.1	0.0039						
event_type	-0.081	-0.097	-0.018	0.037	0.015	-0.016	-0.0041	-0.014	-0.068	-0.064	-0.11	0.12					
num_topics	0.031	-0.14	0.059	-0.00042	0.22	0.006	-0.035	-0.22	0.049	-0.11	-0.0094	0.046	0.0072				
job1_avg_view	0.64	0.37	0.11	0.04	-0.022	-0.013	-0.01	0.023	0.29	0.31	0.0078	-0.042	-0.047	0.015			
event_wise_avg_views	0.16	0.077	0.095	0.021	-0.16	0.015	-0.044	0.17	0.06	0.16	0.025	0.00026	-0.5	-0.018	0.1		
topic_avg_view	0.46	0.21	0.086	0.016	0.053	-0.035	-0.015	-0.052	0.45	0.17	-0.0023	-0.036	-0.015	-0.0035	0.42	0.052	

- Potential predictors for Views are:
  1. Main speaker’s occupation wise average views
  2. Number of comments
  3. Average topic wise views
  4. Number of languages
  5. Average views of all related talks
  6. Event wise average views
- To remove redundancy, Main speaker’s occupation wise average views has been dropped
- To remove data leakage, Number of comments has been dropped

Basis Correlation Matrix, SelectKBest filter method using f-statistics, Sequential Forward Selection (wrapper) and trends identified during EDA, following 08 predictor variables were selected:

1. Published on Weekday or Weekend
    - Talks published on weekdays have more views per Talk than those published on weekends
2. Number of Languages (Positive correlation)
3. Average of views of all related talks (Positive correlation)
4. Duration of Talk
5. Lifetime, i.e., How many days ago Talk was published
6. Number of Speakers
7. Average views of the organizing Event of the talk
8. Average of views of all topics of the talk (Positive correlation)

Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Feature Extraction & Selection

D. Model Creation and Evaluation

E. Key Takeaways & Conclusion

Before Model Creation

- **Train-Test splitting**: Test size of 20% data was chosen
- **Feature scaling**:
  - MinMaxScaler was used for normalization
  - To avoid data leakage, scaler transformation was firstly fitted on training data and then based on the statistical parameters learned from training data, the same transformation was applied on both train and test dataset
- **Evaluation Metric**:
  - MAE (Mean Absolute Error) is used as metric for evaluating various models
  - Reason being it treats large errors/outliers and small errors the same way and does not heavily penalize outliers as MSE or RMSE (Root Mean Squared Error) does

Baseline Models

S. No.	Regression Models	Train RMSE	Test RMSE	Train MAE	Test MAE	Train adjusted-R2 Score	Test adjusted-R2 Score
1	Linear	823153	793975	626543	599176	0.45	0.42
2	Lasso	823153	793974	626544	599176	0.45	0.42
3	KN Neighbors	704307	818304	514202	602214	0.60	0.39
4	Decision Tree	0	1150856	**0**	**813223**	1.00	-0.21
5	Gradient Boosted machines	706298	764408	526892	567055	0.59	0.47
6	XGBoost	712569	758695	**531416**	**561984**	0.59	0.47
7	RandomForest	301820	778963	**222090**	**582559**	0.93	0.44
8	ExtraTrees	712569	758695	**531416**	**561984**	0.59	0.47

- Linear algorithm is not performing good, since there were no linear relationships between predictors and Target variable
- Decision Tree is overfitting and giving worst test performance
- Xgboost, RandomForest and ExtraTrees seem promising, performing good both train and test data

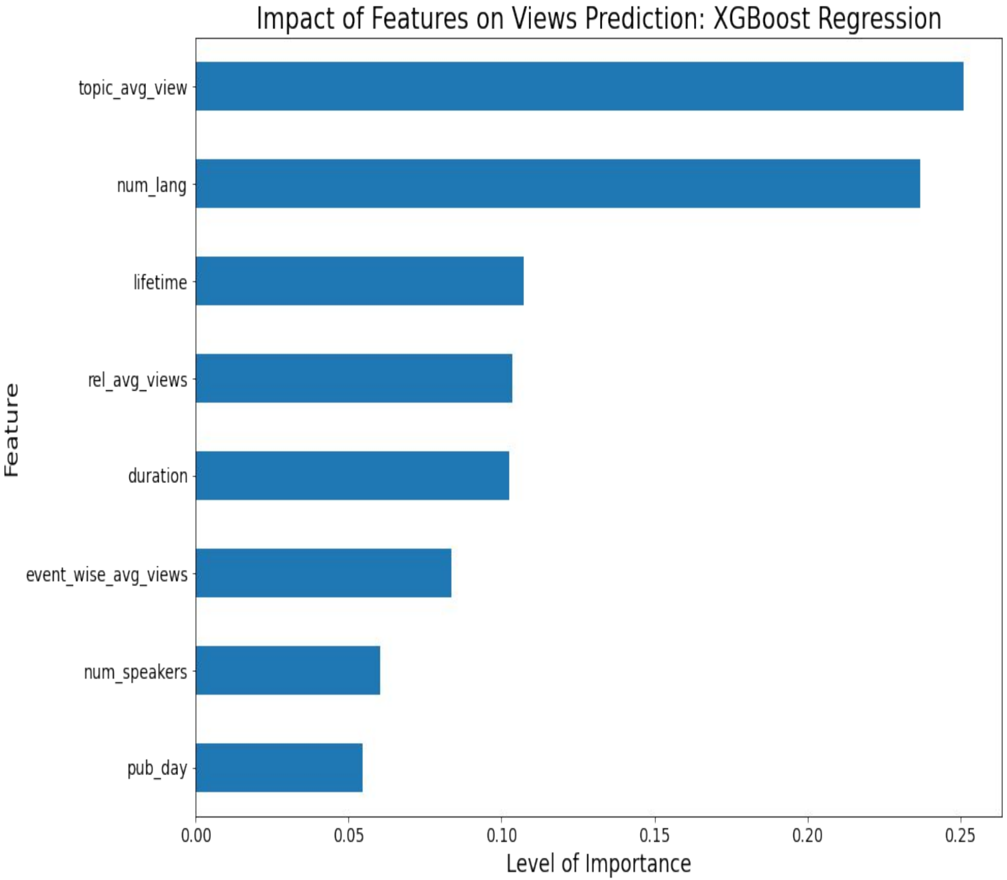
XGBoost: Hyper-parameter Tuning

Optimal Model Parameters:

- eval_metric = 'mae',
- n_estimators = 400,
- learning_rate = 0.02,
- max_depth = 5,
- Gamma = 0.2,
- Lambda = 1,
- min_child_weight = 30,
- Subsample = 0.5

Most Important Features:

1. Average views of all topics,
2. Number of Languages
3. Number of days ago a talk was published



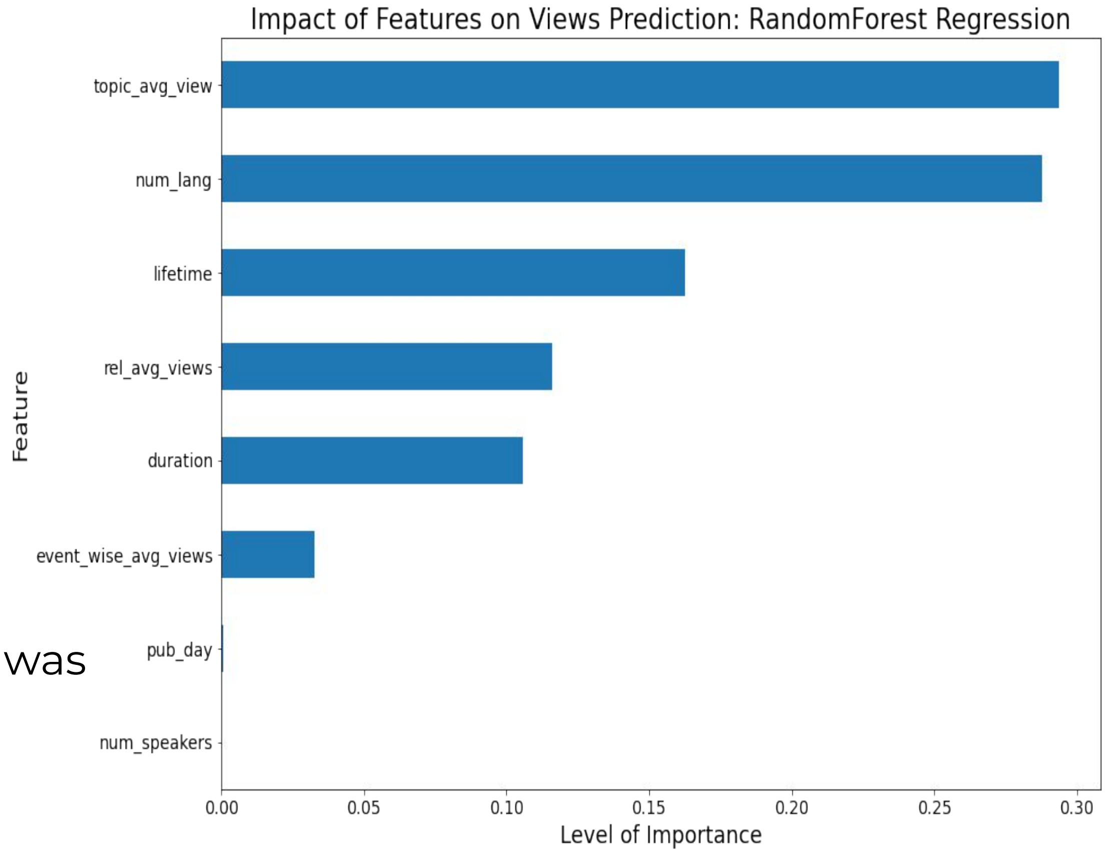
RandomForest: Hyper-parameter Tuning

Optimal Model Parameters:

- Criterion = ‘absolute_error’,
- n_estimators = 400,
- max_depth = 25,
- max_features = 0.5,
- min_samples_leaf = 15,
- min_samples_split = 15,

Most Important Features:

1. Average views of all topics,
2. Number of Languages
3. Number of days ago a talk was published



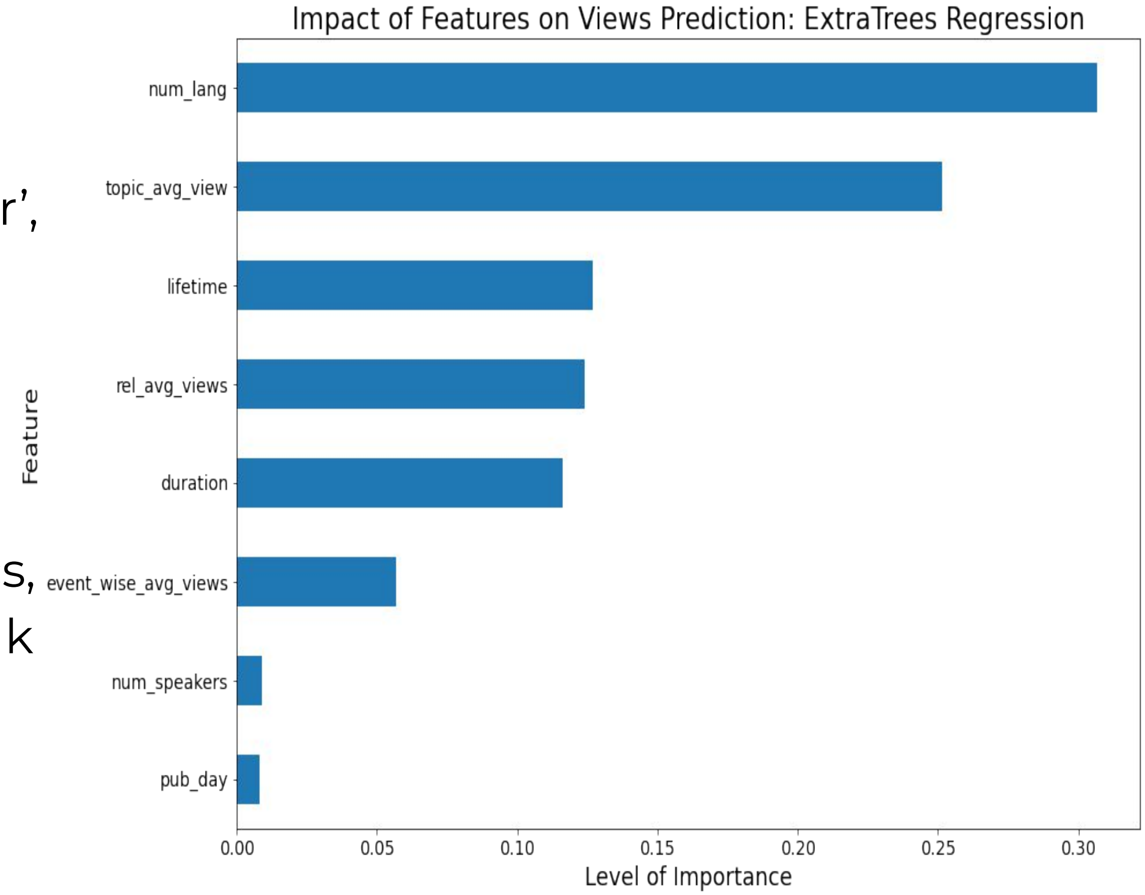
ExtraTrees: Hyper-parameter Tuning

Optimal Model Parameters:

- scoring = ‘neg_mean_absolute_error’,
- n_estimators = 800,
- min_samples_split = 10

Most Important Features:

1. Number of Languages,
2. Average views of all topics,
3. Number of days ago a talk was published



Models After Hyper-parameters Tuning

S. no.	**Regression Models**	**Train RMSE**	**Test RMSE**	**Train MAE**	**Test MAE**	**Improvement in Test MAE after Hyper-parameter Tuning**
1	**XGBoost**	673721	738849	**495211**	**542575**	3.50%
2	RandomForest	739851	784932	515062	561101	3.68%
3	ExtraTrees	460179	767706	337828	572852	1.93%

- **Approach Taken for tuning hyperparameters**:
  - Firstly, I started with random search on the typical recommended values
  - Then did more specific grid search close to the optimal values found in the previous step
- **After Tuning**:
  - Performance of XGBoost and RandomForest models improved significantly, while that of ExtraTrees did not improve much
  - XGBoost is the best performing model with MAE of ~543K views on unseen test dataset

Agenda:

A. Introduction

B. Exploratory Data Analysis and Visualization

C. Feature Extraction & Selection

D. Model Creation and Evaluation

E. Key Takeaways & Conclusion

What can get a Talk Maximum views?

- Talks delivered by writers, authors, psychologist, journalist and neuroscientist
- Talks given by speakers such as Amy Cuddy, James Veitch and Tim Urban
- Talks in native Language English and their availability in more languages
- Talks based on topics such as Creativity, Body Language, Well-being, Success/Productivity, Science and Comedy
- Talks of duration around 20-25 minutes generally have the highest views. Views decreases as talks become either longer or shorter
- Ted talks published on Friday or Wednesday have highest average views, compared to those published on weekends
- Comments drives views and vice-versa

Challenges Faced

- **How to include information contained in non-numerical features into Model creation?**
  1. Temporal Features (Recorded and Published date)
     - Created features like day, month and year of publishing
     - Found that Talks uploaded on Friday have highest average views
  2. Categorical Features (Occupations, Topics, Related Talks and Events)
     - Using target variable, encode above categorical features using mean values
- **Identifying and Eliminating Sources of Data Leakage:**
  1. Comments feature
  2. Mean Target encodings of Categorical features
     - To avoid data leakage, encodings were done using train data only, keeping test data totally unseen by model during training
  3. Scaling/Transformation of features

Conclusion

- Found factors driving views on a given TED Talk
- On average, achieved Test MAE score of ~560K views
- With Hyper-parameter tuning, XGBoost turned out to be the best performing model, with MAE score of ~542K views on test dataset
- Further, to improve model it turns out we may explore other textual features like title, description, transcript as well using NLP. This will be taken upon later in future.

END OF PRESENTATION
THANKS